DØ Data handling, Mass Storage, & Databases

Chip Brock
Michigan State University
for DØ the Data handling/Database/Datagrid Group

CD Practice for

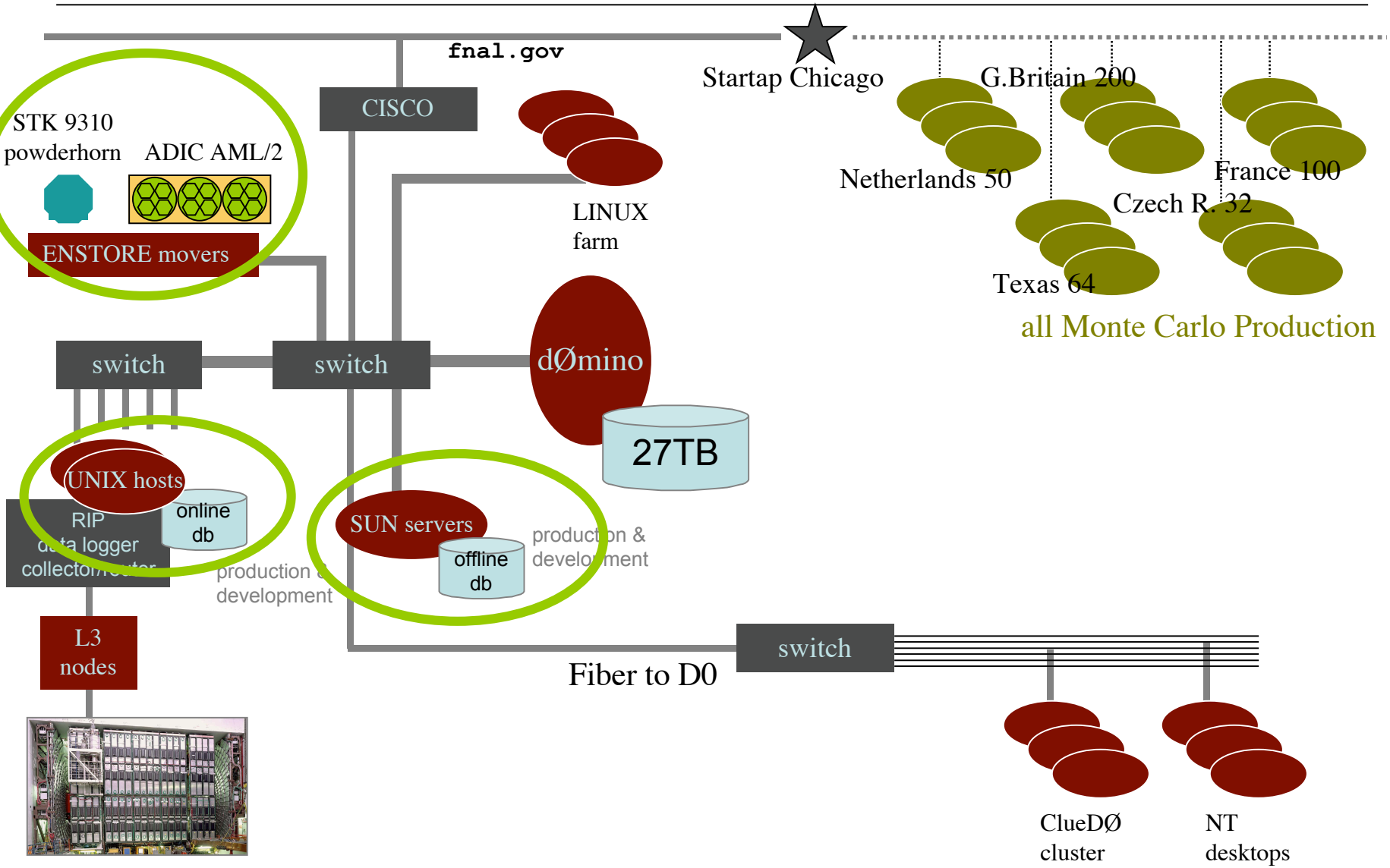
DØ Director's Computing Review

5 June 2002

outline

- **Data handling**
  - overview & philosophy
  - model for planning
- **Mass storage - tape, disk**
  - installed status
  - costs and projections
- **Databases**
  - philosophy
  - installed status
  - costs and projections

DØ data handling/database system architecture



data storage, overview and philosophy

centralized robotic tape handling - utilizing SAM and ENSTORE

- ***storing*** **the data formats…current plans: 5 major tiers** [event size]

specifics under review

1. raw [250KB]
2. raw/RECO "debug" [500KB]
   raw plus the information from reconstruction
3. DST [150KB] - *not yet deployed*
   suitable for much reprocessing, vertices, track clusters, etc.
4. "Thumbnail" (TMB) [10-15KB]
   for event selection and some analysis, multiple vertices, fitted tracks
5. derived set [~10's KB]
   Root tree, ntuple, private format...user defined, needs based

- ***for use by*** **the integrated local and remote systems which service:**
  - online DAQ
  - local production farm
  - local analysis
  - exclusively remote Monte Carlo production
  - hopefully, significant remote analysis

ambitious proposal under review

MSS and data management

ENSTORE for tape-based data storage and delivery

- **STK powderhorn 9310 silos & ADIC AML/2 libraries**
  - Operations support by ISD
    - – managing robotics, drives, tapes, incl. scheduling, maintenance & upgrades
- **dCache (under consideration by DØ)**
  - a means of moderating differing access rates
    - – a performance-enhancing improvement, esp. for off-site access
    - – interface to standard file transfer protocols, like kerberized FTP GridFTP
    - – developed by ISD in collaboration with DESY

SAM is the DØ system for file management

- **layered between the analysis job and MSS…anywhere**
- **manages the file catalog and records file metadata there**
  - maintains record of all stations' access and processing activities
  - provides central logging for debugging and other statistics
- **delivers files from tape to cache for analysis**
  - remembering frequently accessed files
  - intelligently adjudicating tape requests to minimize mounts
  - pre-analyzing resource needs for efficiency

SAM's use in DØ

fully distributed

- **within DØ it currently functions:**
  - between central analysis server and ENSTORE
  - between the reconstruction farm and ENSTORE
  - among the 6 Monte Carlo Farms, Europe and US & >24 active analysis stations worldwide
  - being brought up on the distributed analysis cluster

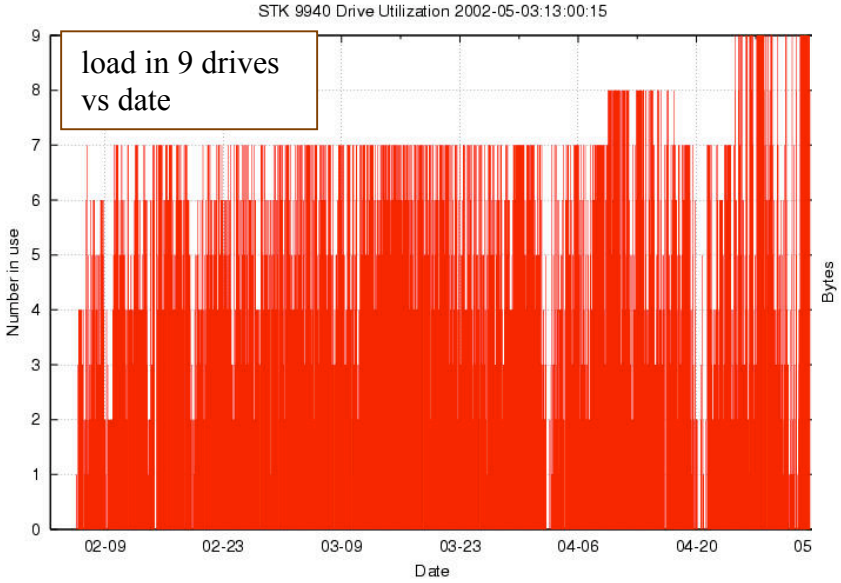
some statistics:

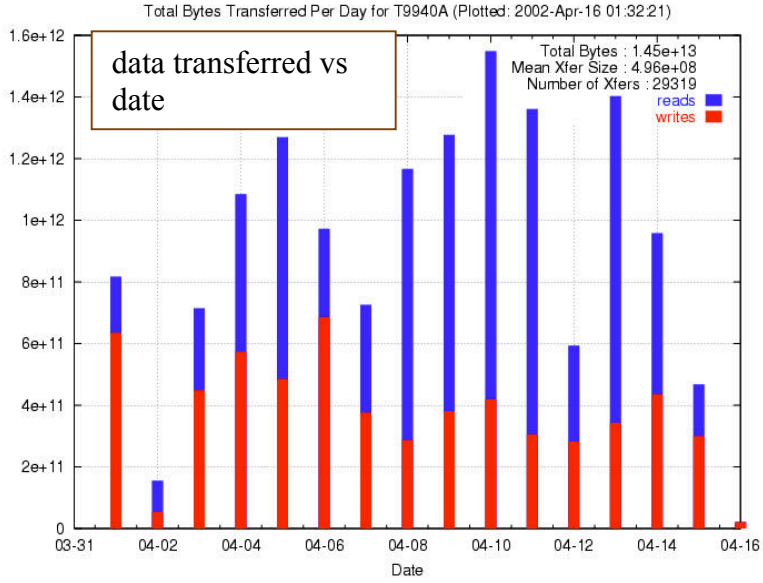
- **496 registered SAM users in production**
  - 334 registered nodes; 48 registered stations, ~24 active
- **57,698 cached files on disk somewhere**
  - 22,242 dataset definitions, 24,075 datasets
  - 65,265 analysis projects
- **651,097 physical and virtual data files**
  - physical: 126,876 raw (on tape); 191,300 reconstructed,
  - 119,905 Root tuple files

MSS: installed capacity

units				current experience		future?
library units	drives R/W rate	drives/capacity slots/unit	cartridge capacities(GB)	tapes written	stored (TB)	
1 STK powderhorn 9310 (data)	9940A 10MB/s	9/20 5500	60	843	48	9940B 120-200GB 20-40MB/s
3 ADIC AML/2 Quadro Towers (MC)	LTO 12MB/s	6/20 3840	100	277	24	IBM/LTO 200-400GB? 20-40MB/s

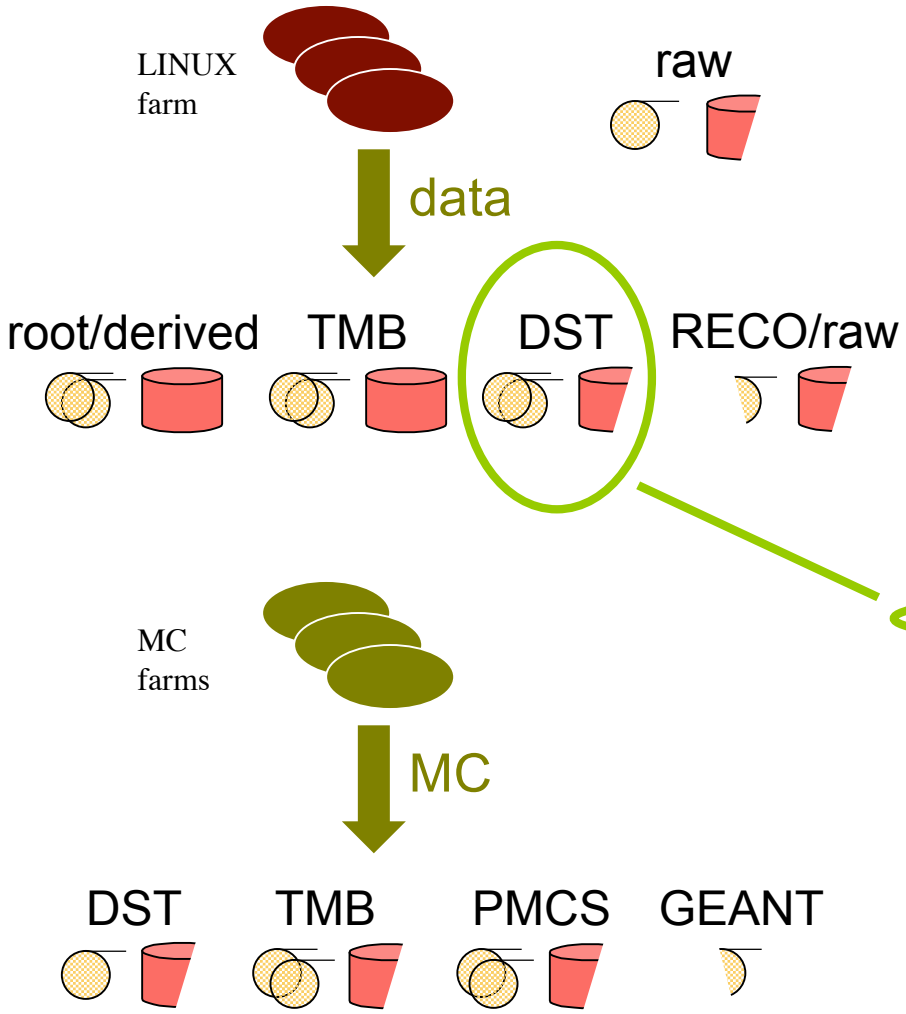
STK experience:





experience has been good…1 bad tape each, files recovered.

model of storage for projections:



In order to control costs, not all data formats will be kept on disk

Modeled by assigning storage in multiples (or fractions) of the number of raw data events:

	size	tape factor	disk factor
raw event	0.25 MB	1	0.001
raw/RECO	0.5 MB	0.2	0.001
data DST	0.15 MB	1.2	0.1
data TMB	0.01 MB	2	1
data root/derived	0.01 MB	8	0
MC D0Gstar	0.7 MB	0.1	0
MC D0Sim	0.3 MB	0	0
MC DST	0.15 MB	1	0.2
MC TMB	0.02 MB	3	0.5
PMCS MC	0.02 MB	2	0.5
MC rootuple	0.02 MB	0	0

With this allocation of the data among tiers and between tape and disk, model the time dependence. Specifically:

storage requirements: tape

Presume 2 running phases:

"Run IIA": **2003-2004**: events as above, with 25Hz mean rate

"Run IIB": **2006-2009**: events 25% larger, with 50Hz mean rate

	1 day	1 year	phase 1 2 years	phase 2 4 years
event rate	2.16E+06	7.88E+08	1.58E+09	6.31E+09
TAPE data accumulation (TB)				
raw event	0.54	197.10	394.20	1971.00
raw/reprocessing	0.22	78.84	157.68	788.40
data DST	0.39	141.91	283.82	1419.12
data TMB	0.04	15.77	31.54	157.68
data rootuple	0.17	63.07	126.14	630.72
MC D0Gstar	0.15	55.19	110.38	551.88
MC D0Sim	0.00	0.00	0.00	0.00
MC DST	0.32	118.26	236.52	1182.60
MC TMB	0.13	47.30	94.61	473.04
PMCS MC	0.09	31.54	63.07	315.36
MC rootuple	0.00	0.00	0.00	0.00
total storage (TB)	**2**	**749**	**1,498**	**7,490**
total storage (PB)	**0.002**	**0.75**	**1.50**	**7.49**
total storage (GB)	**2,052**	**748,980**	**1,497,960**	**7,489,800**

PB of tape storage

storage requirements: disk

3 kinds: **COTS disks as alternative to tape? ✗**

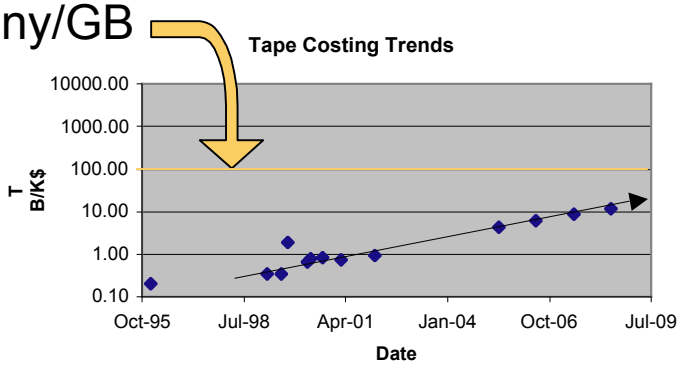
COTS disks to accommodate disk-resident derived data ("data tiers") ~5%

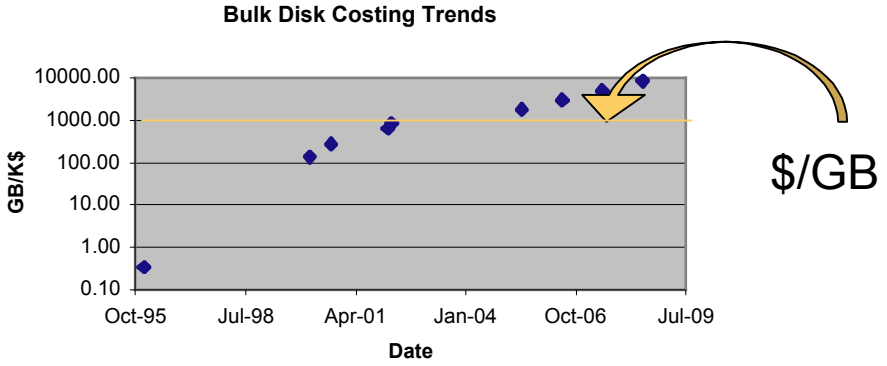
non-COTS disks for database needs

extrapolations from media trends

- **tape + peripherals will likely remain cheaper than COTS disk**

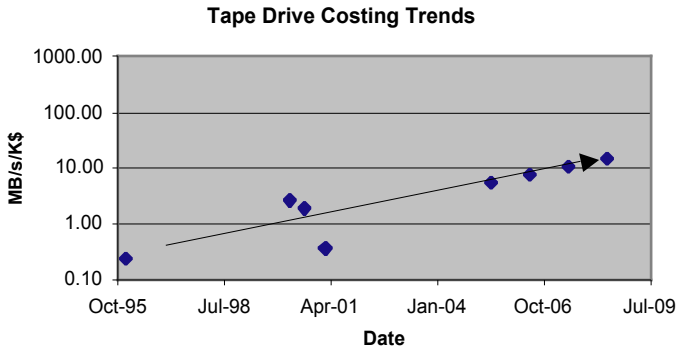
penny/GB





studied extrapolations based on a variety of FNAL experiences plus BTeV, CMS, and CDF predictions

disk costs must also include factor for infrastructure, controllers, cabling, etc.



model, cont.

use extrapolated tape/disk media capacities and prices:

	2003		2005		2007		2009	
	GB	$/GB	GB	$/GB	GB	$/GB	GB	$/GB
STK	120	0.65	250	$0.30	500	0.15	1000	0.07
LTO	200	0.50	400	$0.25	800	0.12	1600	0.06
Disk (COTS)	200	4.00	800	$1.30	3200	0.40	12800	0.25

We have tested many, many scenarios and models.

for example, an STK tape solution:

media costs

	2003	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**TOTALS**
period 1	1	1	0	-1	-1	-1	-1	
period 2	-1	-1	0	1	1	1	1	
STK MB	120,000	120,000	250,000	250,000	500,000	500,000	1,000,000	
$/MB	0.00065	0.000475	0.0003	0.000225	0.00015	0.00011	0.00007	
tapes - p1	6,242	6,242	0	0	0	0	0	12,483
cost - p1	$486,837	$355,766	$0	$0	$0	$0	$0	$842,603
tapes - p2	0	0	0	7,490	7,490	7,490	7,490	29,959
cost - p2	$0	$0	$0	$421,301	$561,735	$411,939	$524,286	$1,919,261
#silo vend4	1.0	2.0	0.0	2.0	2.0	2.0	2.0	11.0
$silo vend4	$75,000	$150,000	$0	$150,000	$150,000	$150,000	$150,000	$825,000

this was done for a variety of possible choices of media, robotics, disk configurations

summarized each iteration as:

TAPE

- **Raw data + data tiers**

media averaged for estimate

A) 100% STK solution

– including additional towers as demanded by tape count

B) 100% ADIC/LTO solution

– including a new trio of QuadroTowers after current towers filled

C) 100% COTS disk solution

DISK

- **Oracle database and SAM disk needs (see later)**

D) includes multiplicative factors for backup, indices, and estimates for drives

next: lots of numbers…

look at the yellow, not the purple

cost projections

STK and LTO tapes differ

competing the falling media costs with rising data rate and robot needs

TOTAL Mass Storage	2003	2004	2005	2006	2007	2008	2009
A) tapes STK	$486,837	$355,766	$0	$421,301	$561,735	$411,939	$524,286
towers+drives	$525,000	$600,000	$600,000	$750,000	$750,000	$750,000	$750,000
total STK alternative	**$1,011,837**	**$955,766**	**$600,000**	**$1,171,301**	**$1,311,735**	**$1,161,939**	**$1,274,286**
B) tapes LTO	$374,490	$280,868	$0	$346,403	$449,388	$337,041	$449,388
mean ADIC libraries + drives	$165,000	$165,000	$220,000	$220,000	$320,000	$320,000	$320,000
total mean ADIC alternative	**$539,490**	**$445,868**	**$220,000**	**$566,403**	**$769,388**	**$657,041**	**$769,388**
C) total data disk alternative	**$2,995,920**	**$1,984,797**	**$0**	**$1,591,583**	**$748,980**	**$608,546**	**$468,113**
D) other disk requirements:							
total tier disk	**$477,770**	**$316,523**	**$0**	**$253,816**	**$119,443**	**$97,047**	**$74,652**
total db/SAM disk	**$52,800**	**$58,035**	**$40,268**	**$34,441**	**$20,259**	**$19,991**	**$18,322**
db/SAM servers	**$60,000**	**$60,000**	**$60,000**	**$60,000**	**$60,000**	**$60,000**	**$60,000**
TOTAL STK alternative	**$1,602,407**	**$1,390,323**	**$700,268**	**$1,519,558**	**$1,511,437**	**$1,338,977**	**$1,427,260**
TOTAL mean ADIC/LTO alternative	**$1,130,060**	**$880,425**	**$320,268**	**$914,660**	**$969,090**	**$834,079**	**$922,362**
TOTAL disk alternative	**$3,586,490**	**$2,419,355**	**$100,268**	**$1,939,839**	**$948,682**	**$785,584**	**$621,086**
TOTAL mean tape media	**$430,664**	**$318,317**	**$0**	**$383,852**	**$505,562**	**$374,490**	**$486,837**
TOTAL tier/db/SAM disk	**$530,570**	**$374,558**	**$40,268**	**$288,256**	**$139,702**	**$117,038**	**$92,974**

mean projected cost of STK and LTO tapes

bottom line: $0.5M/y for tape covers any disappointment in an LTO solution under battlefield conditions

philosophy is : centralized relational databases

online:

calibration (multiple) [P]

runs [P] {config [P], control [P]}

runs quality [D]

hardware [P]

luminosity [p]

offline:

trigger [P] {streaming [d]}

SAM [P]

speakers bureau [P]

run summary [d]

calibration [p]

runs [P] {config [P], control [P], quality [P]}

luminosity [P]

P: production

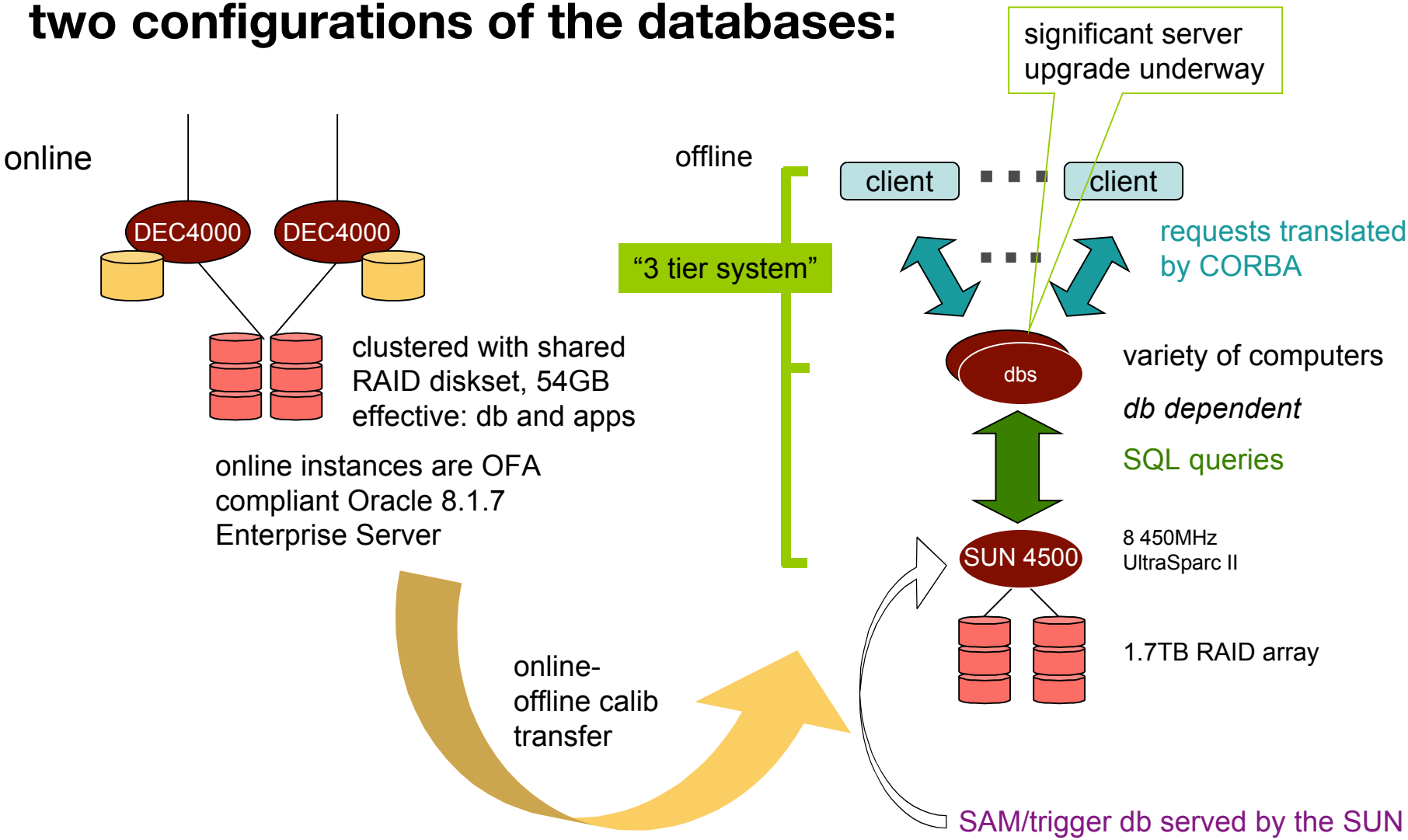
p: barely in production

D: mature development

d: pre-development

application chosen is Oracle at v8.1.7

two configurations of the databases:



db/SAM disk projections

Application Name	Estimated Size 2 Years Run IIa
Offline Calibration Top Level	40 MB
Offline Calibration	90 GB
Offline Muon	30 GB
Offline CFT	14 GB
Offline CPS	2 GB
Offline FPS	8 GB
Offline FPD	small
Offline Luminosity and Streams	200 GB
L1, L2, L3 Trigger	2 GB
SAM File and Event	700 GB
Speakers Bureau	800 MB
VLPC Calibration	7 GB
Run Configuration	105 GB
Total	**1.15 TB**

Assumptions in the next budget tables:

x2 for backup

1.2 for indices

1.2 for development

SAM scales by 1.25/year

(events table scales with $\mathcal{L}$)

look at the yellow, not the purple

db/SAM system projections

TOTAL Mass Storage	2003	2004	2005	2006	2007	2008	2009
A) tapes STK	$486,837	$355,766	$0	$421,301	$561,735	$411,939	$524,286
towers+drives	$525,000	$600,000	$600,000	$750,000	$750,000	$750,000	$750,000
total STK alternative	**$1,011,837**	**$955,766**	**$600,000**	**$1,171,301**	**$1,311,735**	**$1,161,939**	**$1,274,286**
B) tapes LTO	$374,490	$280,868	$0	$346,403	$449,388	$337,041	$449,388
mean ADIC libraries + drives	$165,000	$165,000	$220,000	$220,000	$320,000	$320,000	$320,000
total mean ADIC alternative	**$539,490**	**$445,868**	**$220,000**	**$566,403**	**$769,388**	**$657,041**	**$769,388**
C) total data disk alternative	**$2,995,920**	**$1,984,797**	**$0**	**$1,591,583**	**$748,980**	**$608,546**	**$468,113**
D) other disk requirements:							
total tier disk	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
total db/SAM disk	**$52,800**	**$58,035**	**$40,268**	**$34,441**	**$20,259**	**$19,991**	**$18,322**
db/SAM servers	**$60,000**	**$60,000**	**$60,000**	**$60,000**	**$60,000**	**$60,000**	**$60,000**
TOTAL STK alternative	**$1,602,407**	**$1,390,323**	**$700,268**	**$1,519,558**	**$1,511,437**	**$1,338,977**	**$1,427,260**
TOTAL mean ADIC/LTO alternative	**$1,130,060**	**$880,425**	**$320,268**	**$914,660**	**$969,090**	**$834,079**	**$922,362**
TOTAL disk alternative	**$3,586,490**	**$2,419,355**	**$100,268**	**$1,939,839**	**$948,682**	**$785,584**	**$621,086**
TOTAL mean tape media	**$430,664**	**$318,317**	**$0**	**$383,852**	**$505,562**	**$374,490**	**$486,837**
TOTAL tier/db/SAM disk	**$530,570**	**$374,558**	**$40,268**	**$288,256**	**$139,702**	**$117,038**	**$92,974**

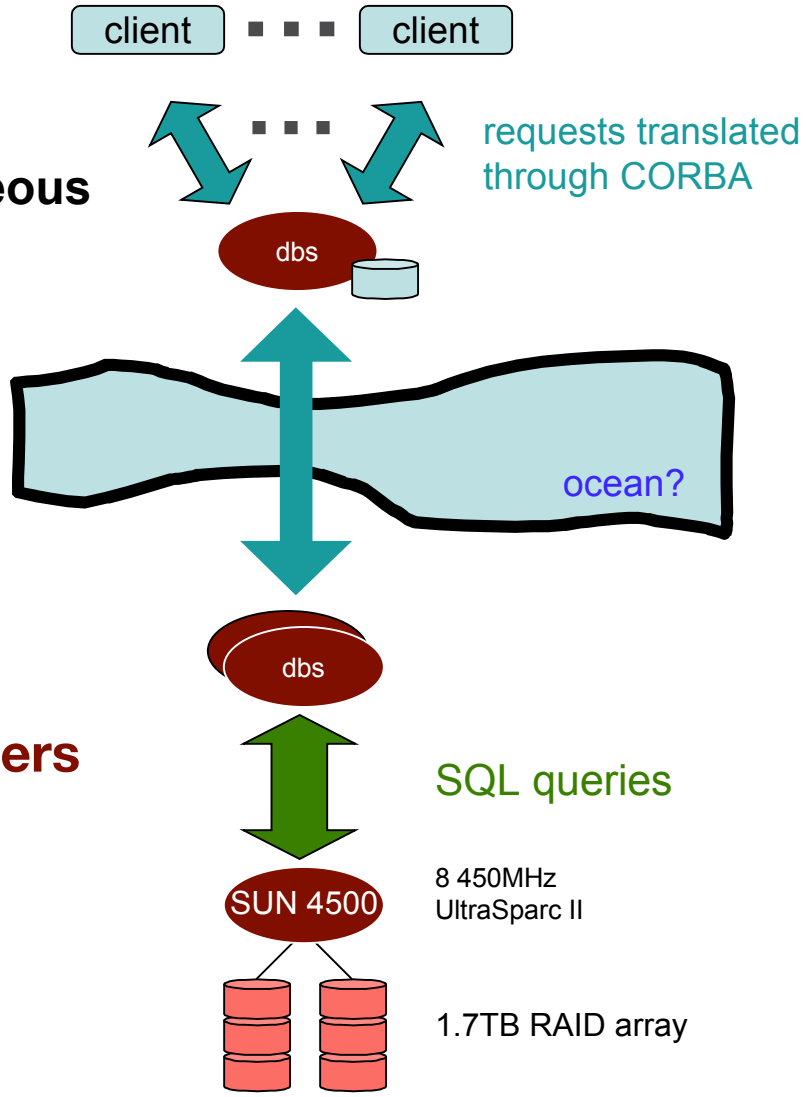
database system plans

server upgrade

- true multi-threading for simultaneous multiple client service
- persistent server side caching
- more monitoring
- more object oriented
- underway now, complete in July

- feature allowing for:

deployment of proxy database servers



datahandling requirements

- **met only with a strategic combination of**
  - lab-based robotic and farm storage
  - lab-based user-provided disk resources
  - off-site/off-shore capabilities and resources

database requirements

- **met with continued**
  - professionals for development & experimentalists for ops
  - nominal hardware increments
  - development of off-site/off-shore database access

we can do it with careful planning